A COMPARISON OF DIETARY PATTERNS DERIVED BY CLUSTER AND PRINCIPAL COMPONENTS ANALYSIS IN A UK COHORT OF CHILDREN

ANDREW D.A.C. SMITH[1] MSTATS

PAULINE M.EMMETT[1] PHD

P. K. NEWBY[2] SCD

KATE NORTHSTONE[1] PHD

[1] School of Social and Community Medicine, University of Bristol, UK

[2] Department of Pediatrics & Program in Graduate Medical Nutrition Sciences, Boston, USA University School of Medicine. Department of Epidemiology, Boston University School of Public Health, USA

RUNNING TITLE: COMPARISON OF DIETARY PATTERN METHODS IN CHILDREN

ADDRESS FOR CORRESPONDENCE

Andrew Smith
School of Social and Community Medicine, University of Bristol
Oakfield House
Oakfield Grove
Bristol
BS8 2BN
UK
Tel: +44 (0)117 33 10104
Email: Andrew.D.Smith@bristol.ac.uk

ABSTRACT

BACKGROUND/OBJECTIVES

To identify dietary patterns in a cohort of 7-year-old children through cluster analysis, compare with patterns derived by principal components analysis (PCA), and investigate associations with socio-demographic variables.

SUBJECTS/METHODS

The main caregivers in the Avon Longitudinal Study of Parents and Children (ALSPAC) recorded dietary intakes of their children (8 279 subjects) using a 94-item food frequency questionnaire. Items were then collapsed into 57 food groups.

Dietary patterns were identified using *k*-means cluster analysis and associations with socio-demographic variables examined using multinomial logistic regression. Clusters were compared with patterns previously derived using PCA.

RESULTS

Three distinct clusters were derived: Processed (4 177 subjects), associated with higher consumption of processed foods and white bread, Plant-based (2 065 subjects), characterized by higher consumption of fruit, vegetables and non-white bread, and Traditional British (2 037 subjects), associated with higher consumption of meat, vegetables and full-fat milk. Membership of the Processed cluster was positively associated with girls, younger mothers, snacking, and older siblings. Membership of the Plant-based cluster was associated with higher educated mothers and vegetarians. The Traditional British cluster was associated with council housing and younger siblings. The three clusters were similar to the three dietary patterns obtained through PCA; each principal component score being higher on average in the corresponding cluster.

CONCLUSIONS

Both cluster analysis and PCA identified three dietary patterns very similar both in the foods associated with them and socio-demographic characteristics. Both methods are useful for deriving meaningful dietary patterns.

KEYWORDS

Dietary patterns, children, cluster analysis, principal components analysis, ALSPAC

INTRODUCTION

Analysis of dietary patterns, as opposed to individual foods or nutrients, has become popular in nutritional epidemiology. Dietary patterns facilitate study of the whole diet, recognizing that people consume foods in combination. They therefore complement traditional methods of examining diet-health relationships that look at individual foods or nutrients.

Most studies that use empirical methods to derive dietary patterns employ principal components analysis (PCA) or cluster analysis (Kant, 2004; Newby and Tucker, 2004). PCA utilises correlations that exist between different food groups, identifying linear combinations of foods that are frequently consumed together. Each subject has a score for every component. Cluster analysis groups individuals into non-overlapping groups (clusters) based on similarities between their diets.

Therefore PCA and cluster analysis describe diet in different ways and it is important to understand their differences. Several studies (Kant et al, 2004; Reedy et al, 2010) have used both methods to investigate associations between dietary patterns and other variables, while other studies (Costacou et al, 2003; Newby et al, 2004; Bamia et al, 2005; Crozier et al, 2006; Hearty and Gibney, 2009) have directly compared the two methods. These studies found some agreement between both methods: all found large differences in mean principal component scores between clusters, particularly for the principal component with the highest variance. We are aware of only one study in children that examined dietary patterns using both methods (Lee et al, 2007).

This study aimed to directly compare dietary patterns obtained in 7-year-old children using cluster analysis and PCA, and examine associations with socio-demographic and lifestyle variables.

METHODS

The Avon Longitudinal Study of Parents and Children (ALSPAC) is an ongoing, population-based study of the environmental and genetic determinants of development and health (Golding et al, 2001). Eligible participants were pregnant women resident in the Avon health authority (South-West England) expected to give birth between 1 April 1991 and 31 December 1992. The core ALSPAC sample included 14 541 women. Data collection was primarily via self-completed questionnaires. Detailed information is available on the ALSPAC website (http://www.bristol.ac.uk/alspac). Ethical approval was granted by the ALSPAC Law and Ethics Committee and Local Research Ethics Committees.

The questionnaire sent to the mother-figure when the child was 81 months old included a food frequency questionnaire (FFQ) on the frequency of consumption of 94 foods and drinks. The questionnaire focused on meals, foods and drinks provided by the mother, so food obtained outside the home (e.g. at school, parties etc.) was not included.

The FFQ measured most of the food frequencies on a five point ordinal scale. These were converted to weekly frequencies of consumption as follows: Never or rarely = 0, Once in 2 weeks = 0.5, 1-3 times a week = 2, 4-7 times a week = 5.5, More than once a day = 10. Other food items (slices of bread, cups of tea and coffee) were continuous measurements. Some items in the FFQ were grouped for ease of interpretation (e.g. shellfish, tuna, uncoated white fish and other fish) resulting in 57 food group variables available for analysis. Since the highest frequency of consumption of each food item was limited, no frequencies were considered to be outliers. Subjects with more than 10 missing values were excluded; otherwise they were assumed to be non-consumers of the missing foods and the frequencies recoded to 0.

Several (self-reported) socio-demographic and lifestyle variables were considered to be potentially associated with dietary patterns. At 81 months the mothers were asked whether they had difficulty getting the child to eat certain foods, whether the child had snacks and/or meals, whether the child was vegetarian, and whether the child had siblings. The mothers reported their own vegetarianism in a questionnaire administered when the child was 47 months old. At 85 months the mothers were asked about their home and lifestyle including whether they lived with a partner, housing tenure (owned/mortgaged, council, rented/other), whether they worked, whether they had difficulties affording food, and their smoking habits. The mother's age and highest educational attainment, whether this was a single or multiple birth, and the child's ethnicity were collected during pregnancy. The child's gender was recorded at birth.

STATISTICAL METHODS

Most cluster analysis algorithms group subjects based on measures of dissimilarity between pairs of subjects. In *k*-means clustering, the main cluster analysis method used in the dietary literature (Newby and Tucker, 2004), subjects are partitioned into clusters that minimize the sum of squares of distances from each subject to the cluster mean. The standard *k*-means algorithm may not always find the cluster solution that minimizes this sum of squares (Everitt et al, 2001, p.100). To mitigate this problem, we ran the algorithm with 100 different starting clusters and chose the solution with the smallest sum of squares.

Input variables are often standardized by subtracting the mean and dividing by the standard deviation. There are potential drawbacks of using this method of standardization for cluster analysis (Everitt et al, 2001, p.51). A more appropriate standardization (Gnanadesikan et al, 1995) is to subtract the mean and divide by the range. We employed this method as it gave cluster solutions that typically explained twice as much of the variation in the sample as those given by the usual method.

Cluster methods are sensitive to small changes in the sample, potentially making solutions unreliable. To test the reliability of the cluster solutions, the data were randomly split into halves and separate analyses performed on each half. This procedure was repeated five times and the number of children allocated to a different cluster was recorded. A further way to assess reliability linear discriminant analysis (LDA). Some studies performed separate analysis on males and females (Pryer et al, 2000; Corrêa Leite et al, 2003; Samieri, 2008; James, 2009; Pryer and Rogers, 2009; Reedy et al, 2010) and this was also assessed.

Analyses were run for 2 to 8 clusters. The best cluster solution was chosen based on the amount of variation in the sample explained by the clusters, size and ease of interpretation of the clusters, and reliability of the solution. Analysis of variance (ANOVA) and the Tukey-Kramer method were used to test for differences between cluster means for each food item.

Associations between clusters and socio-demographic variables were examined using multinomial logistic regression. This gives relative risk ratios, which are the multiplicative changes in the ratio of the probability of membership of one cluster to the probability of membership of another.

RESULTS

Of the 8 515 questionnaires returned, 8 279 (97.2%) were included in the cluster analysis (sufficiently few missing values), of which 6 056 (71.1%) had complete socio-demographic data. Table 1 compares the characteristics, recorded during pregnancy and at birth, of children included in the analysis with those excluded. Children included were more likely to have mothers at least 26 years old at birth, have more educated mothers, live in a mortgaged or owner-occupied home, and have mothers who never smoked (all $p < 0.001$).

A solution with 3 clusters, explaining 14.5% of the variation in the sample, was chosen as the best representation of dietary patterns. In reliability testing at most 32 children (0.4%) were allocated to a different cluster, and LDA resulted in a reclassification of 27 children (0.3%). In this sample, 7 girls changed cluster when boys were removed from the analysis and 11 boys changed cluster when girls were removed, therefore stratification was not performed. The next best solution, with 4 clusters, explained 18.1% of the sample variation but up to 252 children (3.0%) were reallocated in reliability testing.

Table 2 presents mean frequencies of consumption of each food item within each cluster. The largest cluster (4 177 subjects) contained children that consumed, on average, the most white bread, processed meat, fizzy drinks, squash, and snack foods such as chips, crisps, biscuits, sweets and ice cream. Individuals in this cluster consumed less fruit and vegetables, potatoes, bran and oat-based cereals, and water, compared with the other two clusters. We labelled this the *Processed* food cluster.

The remaining two clusters were smaller and of similar size (2 065 and 2 037 subjects). The first contained children that consumed mainly brown/wholemeal bread. Individuals in this cluster consumed more fruit and vegetables, meat substitutes, vegetarian foods, fish, pasta, rice,

bran and oat based cereals, but less meat, fizzy drinks, squash, tea and coffee, and snack foods, compared with the other two clusters. This cluster was named the *Plant-based* cluster.

The third cluster contained children that consumed, on average, the most full-fat milk, and hardly any other milk. Individuals in this cluster consumed, on average, more fruit and vegetables, potatoes, bran and oat based cereals, and water than the Processed cluster, but not as much as the Plant-based cluster. In addition this cluster had the highest consumption of meat and potatoes, and the least consumption of meat substitutes and vegetarian foods. Dairy and puddings were also frequently consumed, with more milk, yoghurt, cakes, buns and puddings than any other cluster, on average. We named this the *Traditional British* cluster.

Table 3 presents associations between clusters and socio-demographic and lifestyle variables. The relative risk ratios presented are shown for each pair of clusters, after adjusting for all other variables. Singleton/multiple birth, child's ethnicity, maternal smoking and employment, whether the mother had a partner, and whether the mother had difficulties getting the child to eat certain foods were not associated with cluster membership (all $p > 0.2$). Maternal age and education were important indicators (both $p < 0.001$). As the mothers' age and education increased the children were much less likely to belong to the Processed cluster.

There was a positive association between vegetarian children ($p = 0.130$), and children with vegetarian mothers ($p = 0.014$), and membership of the Plant-based cluster. Vegetarian children were 1.67 times more likely to belong to the Plant-based cluster than the Processed cluster, and 1.66 times more likely to belong to the Plant-based cluster than the Traditional British cluster. The number of siblings was also associated with cluster membership (both $p < 0.001$). Children with more older siblings were more likely to belong to the Processed cluster, while children with younger siblings were more likely to belong to the Traditional British cluster. Girls and children eating snacks were more likely to belong to the Processed cluster ($p = 0.002$ and $p = 0.003$ respectively). There was some evidence ($p = 0.167$) that children with mothers who had difficulty affording food were also more likely to be in the Processed cluster. Children who were

living in council-housing were more likely to belong to the Traditional British cluster (p = 0.003).

Dietary patterns obtained using PCA on the same variables have been described elsewhere (Northstone and Emmett, 2005). Briefly, three principal components, explaining 18.2% of the variation in the sample, were identified: a 'Junk' pattern with high factor loadings on high-fat, high-sugar, processed and snack foods, a 'Traditional' pattern with high loadings on meat, potatoes and vegetables, and a 'Health conscious' pattern with high loadings on fish, fruit, vegetables, salad and vegetarian style foods. Strong associations between the principal component scores and the above socio-demographic variables have been presented elsewhere (Northstone and Emmett, 2005).

Figure 1 compares the principal component scores (Northstone and Emmett, 2008) with the clusters derived in this study and shows associations between principal components and clusters. The components have been standardized so that they may be compared on the same scale. The `Junk' component score is high in the Processed (mean: 0.19, 95% CI: (0.17,0.22)) and Traditional British (mean: 0.17, 95% CI: (0.13,0.22)) clusters, and lowest in the Plant-based cluster (mean: -0.57, 95% CI: (-0.60,-0.53)). The 'Health-conscious' component is highest in the Plant-based cluster (mean: 0.67, 95% CI: (0.62,0.71)), lowest in the Processed cluster (mean: -0.28, 95% CI: (-0.31,-0.26)), and negative in the Traditional British cluster (mean: -0.09, 95% CI: (-0.13,-0.05)). The 'Traditional' component is highest in the Traditional British cluster (mean: 0.12, 95% CI: (0.06,0.16)), lowest in the Processed cluster (mean: -0.08, 95% CI: (-0.11,-0.05)), and close to zero in the Plant-based cluster (mean: 0.05, 95% CI: (0.00,0.09)).

DISCUSSION

Three clusters were identified in this sample of 7-year-old children: Processed, with high consumption of white bread, processed and snack foods, Plant-based, with high consumption of non-white bread, fruit, vegetables and vegetarian foods, and Traditional British, with high consumption of full-fat milk, meat, potatoes and vegetables.

These clusters generally agreed with the patterns found using PCA in the same sample (Northstone and Emmett, 2005). The 'Junk' factor had high positive loadings for processed and snack foods, and white bread, and a high negative loading on other bread, describing a dietary pattern similar to that described by the Processed cluster. The 'Health conscious' factor had high positive loadings for fruit and vegetables, vegetarian foods, and non-white bread, which are consumed more frequently by the Plant-based cluster. The 'Traditional' factor had high positive loadings for meat, potatoes, vegetables and puddings, but not full-fat milk. Therefore there is some agreement with the Traditional British cluster. The mean factor scores in Figure 1 give further evidence for these agreements.

There were strong associations between clusters and socio-demographic variables. These associations generally agreed with those seen with PCA (Northstone and Emmett, 2005). One difference was girls were positively associated with the Processed cluster but negatively associated with the 'Junk' factor. Another difference was council-housed children were positively associated with the Traditional British cluster but negatively associated with the 'Traditional' factor. Furthermore, difficulties getting the child to eat certain foods, mothers' employment and living with a partner had a stronger association with the factors than the clusters. We found a strong association between the number of older siblings and the Processed cluster, as with the 'Junk' factor. This may be due to older siblings pressuring the mothers to provide certain processed foods aimed at children. Northstone and Emmett (2005) did not find the association between the number of younger siblings and the traditional dietary pattern.

The type of bread and milk consumed was important in determining cluster membership and this may be why girls were more likely to be in the Processed than the Plant-based cluster. Although girls consumed less bread than boys overall, slightly more girls than boys consumed white bread (71.8% compared with 70.4%). Similarly, slightly more girls than boys drank other milk (39.9% compared with 39.2%). As white bread and other milk are strong indicators of the Processed cluster, girls were more likely than boys to belong to that cluster. Similar reasons may explain the association between council housing and the Traditional cluster. Council-housed children consumed, relatively, more full-fat milk than children in other housing types and full-fat milk consumption was strongly associated with the Traditional cluster. These explanations demonstrate not only the importance of bread and milk in the cluster solution, but also the importance of separating foods based on nutrient content. The fact that the associations with these variables differed in PCA may be because milk did not load highly in any of the three factors. PCA combines foods that are consumed together into factors, therefore important single variables, such as milk, can sometimes be missed.

Several studies have examined dietary patterns obtained using cluster analysis in children. Rasanen et al (2002) studied the diet of 7-year-olds in Finland, and found 4 clusters, with large differences in bread and milk consumption between clusters, as in our study. Campain et al (2003) identified 6 clusters, in Australian 12 to 13-year-olds, distinguished by sugar and starch consumption. Knol et al (2005) conducted cluster analyses in low-income US children: 2 to 3-year-olds and 4 to 8-year-olds. There were 6 and 7 clusters present in each age group respectively; no clusters were similar to any in this study. Lee et al (2007) studied diet in Koreans between 1 and 19 years old, identifying 3 clusters (Traditional, Westernized-fast food and Mixed). The 'Westernized-fast food cluster' includes foods similar to our Processed cluster.

We are aware of three studies in children that reported associations between cluster analysis and socio-demographic variables. Song et al (2005) found 2 clusters (Traditional and Modified) among Korean schoolchildren between 12 and 14 years old, but the only significant association

found was between the Traditional pattern and males. Pryer and Rogers (2009) identified 3 clusters (Convenience, Healthy and Traditional) in the diet of British 1 to 5-year-olds. These patterns closely matched our study results, showing similar associations with socio-demographic variables. For example there was a positive association, for boys, between the 'Healthy' cluster and increasing maternal education, as with our Plant-based cluster.

Some studies (Costacou et al, 2003; Kant et al, 2004; Newby et al, 2004; Bamia et al, 2005; Crozier et al, 2006; Reedy et al, 2010) have directly compared the dietary patterns obtained via PCA and cluster analysis, being careful to use the same input variables for both methods. Associations between the two methods were most obvious in those studies with 2 or 3 clusters. Generally, large differences in factor scores between clusters have been reported, particularly for the factor with the largest variance. Often clusters were associated with high or low scores for a particular factor, as in this study.

The similarities between dietary patterns derived by PCA and cluster analysis led Crozier et al (2006) to suggest that PCA is a "more pragmatic choice" as it yields a continuous, rather than categorical, variable for each subject. However, while the derived patterns agree with each other, in this cohort, there are some features that cluster analysis identifies but PCA does not, particularly in milk and bread consumption. In this study the two methods gave contradictory results in terms of the associations with some socio-demographic variables. Both methods have strengths and limitations (Newby and Tucker, 2004; Michels and Schulze, 2005) and can give complementary insights. It is therefore important to consider both methods when establishing dietary patterns in order to gain a deeper understanding of overall diet in a population, and choosing one method over another should be justified. If both methods give similar patterns, then cluster analysis is preferable if a discrete grouping is desired, and PCA is preferable if a continuous outcome is desired.

We conclude that there were three distinct patterns in the diet of ALSPAC children, and these patterns were evident whether derived by cluster analysis or PCA. There were clear

associations with a variety of different socio-demographic and lifestyle factors, although these occasionally differed according to the method used.

ACKNOWLEDGMENTS

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council, the Wellcome Trust and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and KN will serve as guarantor for the contents of this paper. This research was specifically funded by the World Cancer Research Fund (grant number 2009/23).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Bamia C, Orfanos P, Ferrari K, Overvad H, Hundborg HH, Tjonneland A *et al*. (2005). Dietary patterns among older Europeans: the EPIC-Elderly study. *Br J Nutr* **94**, 100-113.

Campain AC, Morgan MV, Evans RW, Ugoni A, Adams GG, Conn JA *et al*. (2003). Sugar-starch combinations in food and the relationship to dental caries in low-risk adolescents. *Eur J Oral Sci* **111**, 316-325.

Corrêa Leite ML, Nicolosi A, Cristina S, Hauser WA, Pugliese P and Nappi G (2003). Dietary and nutritional patterns in an elderly rural population in Northern and Southern Italy: (I). A cluster analysis of food consumption. *Eur J Clin Nutr* **57,** 1514-1521.

Costacou T, Bamia C, Ferrari P, Riboli E, Trichopoulos D and Trichopoulou A (2003). Tracing the Mediterranean diet through principal components and cluster analyses in the Greek population. *Eur J Clin Nutr* **57**, 1378-1385.

Crozier SR, Robinson SM, Borland SE and Inskip HM (2006). Dietary patterns in the Southampton Women's Survey. *Eur J Clin Nutr* **60**, p1391-1399.

Everitt BS, Landau S and Leese M (2001). *Cluster Analysis*, 4th ed. London: Arnold.

Gnanadesikan R, Kettenring JR and Tsao SL (1995). Weighting and selection of variables for Cluster Analysis. *J Class* **12**, 113-136.

Golding J, Pembrey M, Jones R and the ALSPAC Study Team (2001). ALSPAC – The Avon Longitudinal Study of Parent and Children. I. Study methodology. *Paediatr Perinat Epidemiol* **15**, 74-87.

Hearty AP and Gibney MJ (2009). Comparison of cluster and principal component analysis techniques to derive dietary patterns in Irish adults. *Br J Nutr* **101**, 598-608.

James DC (2009). Cluster analysis defines distinct dietary patterns for African-American men and women. *J Am Diet Assoc* **109**, 255-262.

Kant AK (2004). Dietary patterns and health outcomes. *J Am Diet Assoc* **104**, 615-635.

Kant AK, Graubard BI and Schatzkin A (2004). Dietary patterns predict mortality in a national cohort: the National Health Interview Surveys, 1987 and 1992. *J Nutr* **134**, 1793-1799.

Knol LL, Haughton B, Fitzhugh EC (2005). Dietary patterns of young, low-income US children. *J Am Diet Assoc* **105**, 1765-1773.

Lee JW, Hwang J and Cho HS (2007). Dietary patterns of children and adolescents analyzed from 2001 Korea National Health and Nutrition Survey. *Nutr Res Pract* **1**, p84-88.

Michels KB and Schulze MB (2005). Can dietary patterns help us detect diet-disease associations? *Nutr Res Rev* **18**, p241-248.

Newby PK, Muller D and Tucker KL (2004). Associations of empirically derived eating patterns with plasma lipid biomarkers: a comparison of factor and cluster analysis methods. *Am J Clin Nutr* **80**, p759-767.

Newby PK and Tucker KL (2004). Empirically derived eating patterns using factor or cluster analysis: a review. *Nutr Rev* **62**, p177-203.

Northstone K, Emmett P and The ALSPAC Study Team (2005). Multivariate analysis of diet in children at four and seven years of age and associations with socio-demographic characteristics. *Eur J Clin Nutr* **59**, 751-760.

Pryer JA, Cook A and Shetty P (2000). Identification of groups who report similar patterns of diet among a representative national sample of British adults aged 65 years of age or more. *Public Health Nutr* **4**, 787-795.

Pryer JA and Rogers S (2009). Dietary patterns among a national sample of British children aged 1 1/2-4 1/2 years. *Public Health Nutr* **12**, 957-966.

Rasanen M, Lehtinen JC, Niinikoski H, Keskinen S, Ruottinen S, Salminen M *et al*. (2002). Dietary patterns and nutrient intakes of 7-year-old children taking part in an atherosclerosis prevention project in Finland. *J Am Diet Assoc* **102**, p518-524.

Reedy J, Wirfalt E, Flood A, Mitrou PN, Krebs-Smith SM, Kipnis V *et al*. (2010). Comparing 3 dietary pattern methods–cluster analysis, factor analysis, and index analysis–with colorectal cancer risk: The NIH-AARP Diet and Health Study. *Am J Epidemiol* **171**, 479-487.

Samieri C, Jutand MA, Feart C, Capuron L, Letenneur L and Barberger-Gateau P (2008). Dietary patterns derived by hybrid clustering method in older people: association with cognition, mood, and self-rated health. *J Am Diet Assoc* **108**, p1461-1471.

Song Y, Joung H, Engelhardt K, Yoo SY and Paik HY (2005). Traditional v. modified dietary patterns and their influence on adolescents' nutritional profile. *Br J Nutr* **93**, p943-949.

Table 1: Differences in characteristics (collected during pregnancy and at birth) between children with and without dietary data.

		With dietary data		Without dietary data	
Sex					
Girl	(48.3%)	4 019	(48.5%)	2 808	(47.9%)
Boy	(51.7%)	4 260	(51.5%)	3 059	(52.1%)
$\chi^2 = 0.64$ ($p = 0.423$)					
Maternal age					
<21	(7.2%)	299	(3.6%)	712	(12.3%)
21-25	(23.5%)	1 513	(18.3%)	1 793	(30.9%)
26-30	(39.3%)	3 506	(42.4%)	2 027	(34.9%)
31+	(30.1%)	2 961	(35.8%)	1 271	(21.9%)
$\chi^2 = 853.70$ ($p < 0.001$)					
Maternal education					
CSE	(20.2%)	1 184	(14.7%)	1 338	(30.3%)
Vocational	(9.9%)	708	(8.8%)	520	(11.8%)
O level	(34.6%)	2 856	(35.4%)	1 467	(33.2%)
A level	(22.5%)	2 054	(25.5%)	749	(16.9%)
Degree	(12.9%)	1 258	(15.6%)	349	(7.9%)
$\chi^2 = 597.26$ ($p < 0.001$)					
Housing tenure					
Mortgaged/owned	(73.4%)	6 578	(81.6%)	3 294	(61.2%)
Council	(16.0%)	818	(10.1%)	1 335	(24.8%)
Rented/other	(10.6%)	670	(8.3%)	753	(14.0%)
$\chi^2 = 714.20$ ($p < 0.001$)					
Maternal smoking					
Never	(49.2%)	4 322	(53.3%)	2 170	(42.6%)
Ever	(50.8%)	3 790	(46.7%)	2 920	(57.4%)
$\chi^2 = 141.84$ ($p < 0.001$)					

Table 2: Mean (standard deviation) frequency of weekly consumption of foods across cluster for 8 281 children.

Cluster		Processed	Plant-based	Traditional British	F
Food item	Cluster size	4 177	2 065	2 037	
Full-fat milk		0.26 (0.64) [a]	1.08 (2.11) [b]	**7.73 (2.25)** [c]	15 709.50 ($p < 0.001$)
Non-white bread (slices)		0.01 (0.27) [a]	**15.82 (4.41)** [b]	2.33 (5.65) [c]	14 067.43 ($p < 0.001$)
White bread (slices)		**15.76 (4.32)** [a]	0.00 (0.15) [b]	13.35 (7.03) [c]	8 225.01 ($p < 0.001$)
Other milk		2.81 (3.53) [a]	**2.86 (3.65)** [a]	0.00 (0.00) [b]	638.44 ($p < 0.001$)
Salad		2.49 (2.90) [a]	**3.45 (3.32)** [b]	2.69 (3.19) [c]	67.35 ($p < 0.001$)
Fizzy drinks		**3.38 (3.74)** [a]	2.34 (3.01) [b]	3.19 (3.94) [a]	58.70 ($p < 0.001$)
Pasta		1.49 (1.33) [a]	**1.86 (1.35)** [b]	1.51 (1.31) [a]	58.28 ($p < 0.001$)
Biscuits		**6.39 (3.68)** [a]	5.41 (3.50) [b]	6.30 (3.82) [a]	52.67 ($p < 0.001$)
Chips		**2.56 (1.77)** [a]	2.10 (1.69) [b]	2.56 (1.86) [a]	52.37 ($p < 0.001$)
Sausages, burgers		**1.13 (0.96)** [a]	0.89 (0.85) [b]	1.12 (1.00) [a]	50.92 ($p < 0.001$)
Water		3.39 (3.56) [a]	**4.20 (3.64)** [b]	4.17 (3.84) [b]	49.26 ($p < 0.001$)
Crisps		**3.82 (2.27)** [a]	3.22 (2.18) [b]	3.69 (2.34) [a]	48.85 ($p < 0.001$)
Pulses		0.10 (0.49) [a]	**0.25 (0.68)** [b]	0.15 (0.59) [c]	47.50 ($p < 0.001$)
Chicken/turkey in crispy coating		**1.83 (1.33)** [a]	1.50 (1.30) [b]	1.79 (1.39) [a]	44.75 ($p < 0.001$)
Green vegetables		4.28 (3.18) [a]	**5.09 (3.39)** [b]	4.75 (3.53) [c]	44.04 ($p < 0.001$)
Meat substitutes		0.13 (0.57) [a]	**0.28 (0.76)** [b]	0.13 (0.57) [a]	43.67 ($p < 0.001$)
Roast potatoes		1.37 (0.97) [a]	1.15 (0.96) [b]	**1.39 (1.06)** [a]	40.71 ($p < 0.001$)
Sweets		**1.88 (1.68)** [a]	1.49 (1.45) [b]	1.83 (1.76) [a]	40.08 ($p < 0.001$)
Chocolate		2.14 (1.79) [a]	1.76 (1.58) [b]	**2.15 (1.81)** [a]	37.75 ($p < 0.001$)
Fish		1.29 (1.74) [a]	**1.70 (1.86)** [b]	1.37 (1.87) [a]	37.09 ($p < 0.001$)
Bran-based cereal		1.98 (2.05) [a]	**2.44 (2.19)** [b]	2.25 (2.19) [c]	36.30 ($p < 0.001$)
Oat-based cereal		0.70 (1.32) [a]	**1.02 (1.58)** [b]	0.86 (1.48) [c]	35.13 ($p < 0.001$)
Fresh fruit		6.66 (4.65) [a]	**7.71 (4.69)** [b]	7.02 (4.86) [c]	34.34 ($p < 0.001$)
Ice lollies		**1.36 (1.59)** [a]	1.04 (1.30) [b]	1.32 (1.61) [a]	32.84 ($p < 0.001$)
Cheese		2.44 (2.11) [a]	**2.86 (2.15)** [b]	2.74 (2.16) [b]	31.84 ($p < 0.001$)
Meat pies/pasties		0.50 (0.81) [a]	0.36 (0.75) [b]	**0.55 (0.85)** [a]	30.73 ($p < 0.001$)
Nuts		0.68 (1.49) [a]	**1.00 (1.78)** [b]	0.71 (1.46) [a]	29.66 ($p < 0.001$)
Baked beans/tinned pasta		2.69 (1.94) [a]	2.34 (1.77) [b]	**2.70 (1.98)** [a]	26.86 ($p < 0.001$)

Meat dishes	1.97 (1.46) [a]	1.72 (1.36) [b]	**2.01 (1.48)** [a]	26.52 ($p < 0.001$)
Root vegetables	2.82 (2.12) [a]	**3.24 (2.32)** [b]	3.05 (2.31) [c]	26.43 ($p < 0.001$)
Tea/coffee	3.61 (5.81) [a]	2.70 (5.30) [b]	**3.86 (6.21)** [a]	23.83 ($p < 0.001$)
Ice cream	**1.86 (1.60)** [a]	1.59 (1.36) [b]	1.86 (1.65) [a]	23.29 ($p < 0.001$)
Rice	1.02 (1.14) [a]	**1.22 (1.07)** [b]	1.06 (1.17) [a]	22.59 ($p < 0.001$)
Sweetcorn	1.11 (1.22) [a]	**1.32 (1.29)** [b]	1.20 (1.30) [c]	20.58 ($p < 0.001$)
Fruit juice	3.12 (3.17) [a]	**3.65 (3.26)** [b]	3.32 (3.30) [c]	18.88 ($p < 0.001$)
Vegetarian pies/pasties	0.22 (0.66) [a]	**0.30 (0.70)** [b]	0.19 (0.58) [a]	17.01 ($p < 0.001$)
Other cereal	3.38 (2.21) [a]	3.13 (2.15) [b]	**3.52 (2.25)** [a]	16.24 ($p < 0.001$)
Flavoured milk drinks	1.17 (2.02) [a]	1.17 (2.00) [a]	**1.48 (2.48)** [b]	15.89 ($p < 0.001$)
Squash	**6.12 (3.36)** [a]	5.62 (3.49) [b]	6.02 (3.41) [a]	15.14 ($p < 0.001$)
Eggs	1.03 (1.16) [a]	**1.19 (1.14)** [b]	1.13 (1.22) [b]	14.23 ($p < 0.001$)
Cold meats	1.97 (1.81) [a]	1.77 (1.71) [b]	**2.04 (1.87)** [a]	13.28 ($p < 0.001$)
Potatoes	2.11 (1.50) [a]	**2.26 (1.54)** [b]	2.29 (1.62) [b]	12.14 ($p < 0.001$)
Peas	1.26 (1.26) [a]	**1.42 (1.28)** [b]	1.36 (1.34) [b]	11.72 ($p < 0.001$)
Fried food	0.48 (0.78) [a]	0.40 (0.71) [b]	**0.50 (0.82)** [a]	11.51 ($p < 0.001$)
White fish in breadcrumbs/batter	1.31 (0.97) [a]	1.27 (1.00) [a]	**1.39 (1.02)** [b]	7.99 ($p < 0.001$)
Milk-based puddings/custard	1.54 (1.82) [a]	1.52 (1.86) [a]	**1.71 (1.99)** [b]	7.04 ($p < 0.001$)
Other puddings	0.83 (1.09) [a]	0.88 (1.09) [ab]	**0.94 (1.23)** [b]	6.96 ($p < 0.001$)
Crispbreads/crackers	0.36 (0.98) [a]	**0.46 (1.07)** [b]	0.40 (1.07) [ab]	6.28 ($p < 0.001$)
Cakes/buns	2.10 (1.80) [a]	2.03 (1.75) [a]	**2.22 (1.90)** [b]	5.81 ($p = 0.001$)
Herbal tea	0.03 (0.34) [a]	**0.05 (0.42)** [b]	0.02 (0.23) [a]	5.47 ($p = 0.001$)
Poultry	**1.91 (1.27)** [a]	1.82 (1.29) [b]	1.88 (1.32) [ab]	3.52 ($p = 0.014$)
Yoghurt/fromage frais	4.05 (2.63) [a]	4.13 (2.53) [ab]	**4.22 (2.62)** [b]	3.06 ($p = 0.027$)
Pizza	1.06 (0.96) [a]	**1.07 (0.95)** [a]	1.03 (0.95) [a]	1.20 ($p = 0.308$)
Butter/margarine (slices)	14.88 (8.81) [a]	14.97 (8.61) [a]	**15.23 (8.20)** [a]	1.09 ($p = 0.354$)
Offal	0.04 (0.37) [a]	0.04 (0.27) [a]	**0.05 (0.39)** [a]	0.71 ($p = 0.547$)
Alcohol	0.07 (0.26) [a]	**0.08 (0.26)** [a]	0.08 (0.26) [a]	0.12 ($p = 0.951$)
Tinned fruit	**0.64 (1.05)** [a]	0.64 (1.06) [a]	0.63 (1.07) [a]	0.09 ($p = 0.965$)

[abc] Where superscripts differ there is a significant difference between cluster means (Tukey-Kramer method)
The highest and lowest mean in each row are bold and underlined respectively.

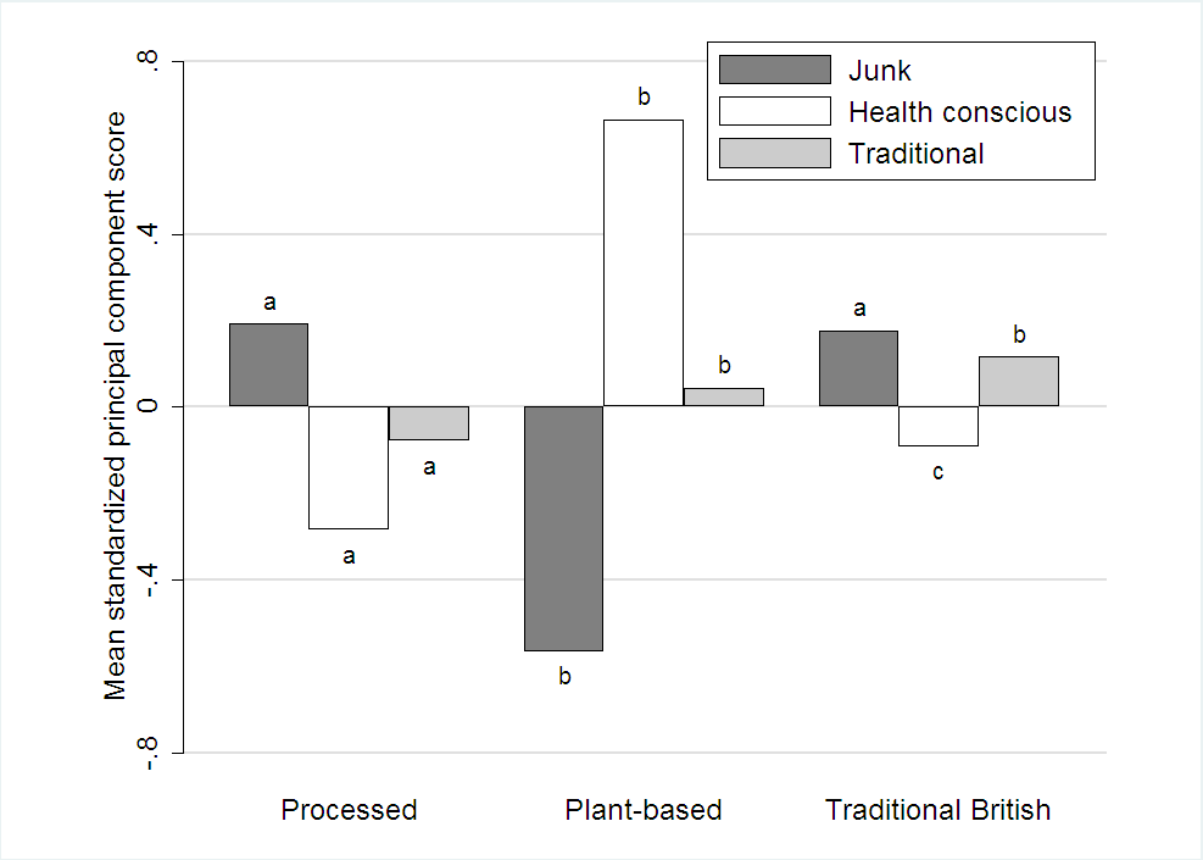
Table 3: Adjusted relative risk ratios (95% confidence intervals) between pairs of clusters, estimated by multinomial logistic regression.

Characteristic	n	Processed Plant-based	Plant-based Traditional British	Traditional British Processed
Children in pregnancy				
Singleton	5 914	1	1	1
Multiple birth	142	1.04 (0.68, 1.61)	0.95 (0.58, 1.56)	1.01 (0.66, 1.54)
$p = 0.975$				
Sex				
Girl	3 115	1	1	1
Boy	2 941	0.82 (0.72, 0.93)	1.03 (0.89, 1.20)	1.18 (1.04, 1.34)
$p = 0.002$				
Ethnicity of child				
White	5 876	1	1	1
Nonwhite	180	0.66 (0.45, 0.96)	1.03 (0.69, 1.54)	1.47 (1.01, 2.13)
$p = 0.263$				
Maternal age				
<21	130	1	1	1
21-25	994	0.59 (0.33, 1.07)	1.07 (0.56, 2.05)	1.57 (1.02, 2.43)
26-30	2 644	0.52 (0.29, 0.92)	1.20 (0.64, 2.28)	1.60 (1.04, 2.46)
31+	2 288	0.37 (0.21, 0.67)	1.50 (0.79, 2.88)	1.77 (1.13, 2.76)
$p < 0.001$				
Maternal education				
CSE	740	1	1	1
Vocational	504	0.84 (0.60, 1.17)	1.19 (0.82, 1.72)	1.01 (0.76, 1.32)
O level	2 163	0.65 (0.51, 0.83)	1.46 (1.10, 1.94)	1.05 (0.86, 1.30)
A level	1 604	0.42 (0.33, 0.55)	2.01 (1.50, 2.69)	1.18 (0.95, 1.48)
Degree	1 045	0.30 (0.23, 0.39)	2.75 (2.00, 3.76)	1.22 (0.94, 1.57)
$p < 0.001$				
Mother works				
Yes	4 220	1	1	1
No	1 836	0.95 (0.82, 1.10)	0.97 (0.82, 1.14)	1.09 (0.94, 1.25)
$p = 1.000$				
Difficulties affording food				
None	5 383	1	1	1
Some	646	1.17 (0.94, 1.47)	0.87 (0.67, 1.12)	0.98 (0.80, 1.20)
Yes	27	3.23 (0.72, 14.7)	0.19 (0.04, 0.89)	1.61 (0.70, 3.65)
$p = 0.167$				
Housing tenure				
Mortgaged/owned	5 396	1	1	1
Council	414	0.95 (0.69, 1.30)	0.64 (0.45, 0.90)	1.66 (1.29, 2.13)
Rented/other	246	0.85 (0.62, 1.17)	1.29 (0.89, 1.87)	0.91 (0.65, 1.27)
$p = 0.003$				
Mother has a partner				
Yes	5 839	1	1	1
No	217	0.94 (0.66, 1.33)	1.51 (0.99, 2.32)	0.70 (0.48, 1.02)
$p = 1.000$				

Maternal smoking				
No	5 011	1	1	1
<10 per day	374	1.07 (0.81, 1.41)	0.89 (0.65, 1.22)	1.05 (0.81, 1.36)
10-19 per day	454	1.08 (0.82, 1.42)	0.78 (0.57, 1.06)	1.19 (0.93, 1.51)
20+ per day	217	1.28 (0.84, 1.93)	0.58 (0.37, 0.90)	1.35 (0.98, 1.87)
$p = 0.375$				
Mother vegetarian				
No	5 747	1	1	1
Yes	309	0.73 (0.53, 1.01)	1.20 (0.84, 1.73)	1.13 (0.80, 1.60)
$p = 0.014$				
Child vegetarian				
No	5 875	1	1	1
Yes	181	0.60 (0.40, 0.90)	1.66 (1.04, 2.64)	1.01 (0.64, 1.59)
$p = 0.130$				
Ch diet based on snacks/meals				
Meals/no snacks	3 657	1	1	1
Snacks and Meals	2 141	1.39 (1.20, 1.60)	0.93 (0.79, 1.10)	0.77 (0.67, 0.89)
Snacks/no meals	46	1.01 (0.45, 2.25)	0.90 (0.37, 2.16)	1.11 (0.55, 2.24)
Other	212	0.92 (0.66, 1.28)	1.28 (0.86, 1.91)	0.85 (0.59, 1.23)
$p = 0.003$				
Child difficult eater				
No	2 683	1	1	1
Yes, sometimes difficult	1 758	1.17 (0.89, 1.54)	0.79 (0.58, 1.07)	1.09 (0.84, 1.40)
Yes, difficult	1 174	1.10 (0.93, 1.32)	0.78 (0.64, 0.96)	1.16 (0.97, 1.37)
Yes, very difficult	441	0.99 (0.85, 1.16)	0.94 (0.79, 1.12)	1.07 (0.92, 1.25)
$p = 0.955$				
Number of older siblings				
None	2 755	1	1	1
1	2 317	1.21 (1.03, 1.42)	1.12 (0.94, 1.36)	0.73 (0.62, 0.86)
2 or more	984	1.58 (1.28, 1.97)	0.99 (0.76, 1.27)	0.64 (0.52, 0.80)
$p < 0.001$				
Number of younger siblings				
None	2 946	1	1	1
1	2 490	1.01 (0.86, 1.19)	0.58 (0.48, 0.71)	1.69 (1.44, 1.99)
2 or more	620	1.21 (0.92, 1.57)	0.43 (0.33, 0.58)	1.90 (1.50, 2.40)
$p < 0.001$				

p values are based on partial sums of squares

Figure 1: Mean standardized principal component scores for each cluster. Both factor scores and clusters were obtained from the same FFQ variables in the same sample of 8 281 children.



[abc] Where letters differ there is a significant difference between cluster means (Tukey-Kramer method)